The communicative import of gesture

Mol, L.; Krahmer, E.J.; Maes, A.; Swerts, M.G.J.

Published in:
Gesture

Publication date:
2009

The communicative import of gestures

Evidence from a comparative analysis of human–human and human–machine interactions

Lisette Mol, Emiel Krahmer, Alfons Maes, and Marc Swerts
Tilburg University

Does gesturing primarily serve speaker internal purposes, or does it mostly facilitate communication, for example by conveying semantic content, or easing social interaction? To address this question, we asked native speakers of Dutch to retell an animated cartoon to a presumed audiovisual summarizer, a presumed addressee in another room (through web cam), or an addressee in the same room, who could either see them and be seen by them or not.
We found that participants produced the least number of gestures when talking to the presumed summarizer. In addition, they produced a smaller proportion of large gestures and almost no pointing gestures. Two perception experiments revealed that observers are sensitive to this difference in gesturing. We conclude that gesture production is not a fully automated speech facilitation process, and that it can convey information about the communicative setting a speaker is in.

Keywords: gesture, human–machine interaction, narration, audience design

Introduction

In this paper we explore the functional roles of spontaneous hand gestures produced during narrative speech, by looking at it from the production as well as the perception perspective. If gesture production primarily serves speaker internal processes, then we would expect it to be a highly automated process that is little influenced by the communicative setting a speaker is in, and by whom a speaker is addressing. On the other hand, if gestures primarily aid communication between a speaker and an addressee, then we would expect gesturing to be a more flexible process, which is adapted to different communicative environments and audience characteristics.

If speakers gesture mostly for themselves, then addressees may or may not be able to use information from gestures. However, if addressees are unable to use

Gesture 9:1 (2009), **97–126**. DOI 10.1075/gest.9.1.04mol
ISSN 1568–1475 / E-ISSN 1569–9773

information from the gesture modality, this would make it less likely that speakers intend their gestures communicatively. Since people continuously switch roles between speaker and addressee in day-to-day communication, we think it unlikely that speakers would put communicative effort into a modality that they never use as addressees. And for the same reason, if speakers gesture partly to communicate, then we would expect that addressees are able to gain information from speakers' gestures.

In this paper we describe two studies. First we describe an experiment from the speaker's perspective, in which we manipulated the nature of the addressee (either artificial or human) and whether speaker and addressee could see each other. We were interested in the effects of these manipulations on gesture production. Our second study consists of two perception experiments using video-clips from the first study. We measured whether addressees were sensitive to possible differences in gesturing that resulted from a speaker addressing a human or an artificial addressee.

Background

The functional role of gestures

Many studies have been conducted to investigate the primary functional role of hand gestures. One view is that gestures are mostly produced for the benefit of the speaker, for example to aid speech production. Hadar (1989), Krauss (1998), de Ruiter (1998), Kita (2000), Hostetter and Hopkins (2002), Hostetter, Alibali and Kita (2007), among others, have found support for this view. Some studies have shown that gesturing may facilitate cognition in processes other than language production, which is another for-speaker function (Goldin-Meadow et al., 2001; Goldin-Meadow, 1999, p. 427). Another view is that speakers produce gestures with a communicative intent. Kendon (2004), for example, argues that speakers produce gestures as an integral part of their communicative effort. Support for this hypothesis has been found, among others, by Cohen and Harrison (1973), Cohen (1977), Özyürek (2002), Jacobs and Garnham (2006), and Bangerter and Chevalley (2007). (Also see Kendon, 1994.)

Alibali, Heath, and Myers (2001) have tried to reconcile various seemingly contradictory experimental results by associating different types of gestures with different functional roles. They conducted a study in which narrators told a story to an addressee either face-to-face, or with a wooden screen in between speaker and addressee. They found that speakers produced more representational gestures (gestures that depict some of the content of the story) in the face-to-face condition

than in the screen condition, when the addressee could not see the speaker, although representational gestures were also produced in this latter condition. Beat gestures (gestures that do not depict narrative content), on the other hand, were produced at comparable rates under both conditions.[1] The fact that speakers still produced many (representational) gestures when it was clear that the addressee could not see them is not easily explained by a theory that stresses the communicative function of gestures. Alibali et al. concluded that both types of gesture serve both speaker-internal and communicative functions. They suggested examining "how different speakers use gestures in different types of contexts for both speaker-internal and communicative purposes" rather than trying to find a single primary role of gesture production. Let us briefly review several factors that have been suggested as having an influence on gesture production, and that are relevant to the present study.

Factors influencing gesture production

Visibility and dialogue. Bangerter and Chevalley (2007) investigated the effect of mutual visibility on pointing gestures in a referential communication task. They found that pointing movements that did not involve raising the arm, were produced at equal rates, regardless of whether conversational partners could see each other or not. This suggests that they are automatic in production. However, pointing movements that did involve raising the arm were used more when interlocutors could see each other, suggesting that they are intended to communicate. Thus, gesture size seems to be indicative of the gesture's functional role, and of the nature of the cognitive processes underlying its production.

In a somewhat similar vein, Enfield, Kita, and de Ruiter (2007) describe a theory of how different sizes of pointing gestures serve different pragmatic functions in face-to-face communication. Based on data from Lao, they argue that larger pointing gestures carry primary, "informationally foregrounded" information, whereas smaller pointing gestures carry "informationally backgrounded information, which refers to a possible but uncertain lack of referential common ground".

The importance of gesture size in relation to visibility was also found by Bavelas et al. (2008). In a picture description task, they compared face-to-face communication (which enables dialogue and visibility) to talking through a hand held phone (dialogue, but no visibility) and talking to a tape recorder using a hand held microphone (no dialogue, no visibility). They found that speakers gestured more while being engaged in dialogue, and also that they gestured very differently if there was the possibility to demonstrate things to the addressee by gesture. Participants described a picture of an old-fashioned dress. In the face-to-face condition, gestures were done to describe features of the dress as if it was full size. In

the phone condition, gestures were only the size of the picture, and were harder to interpret. In the tape recorder condition, gestures were very small and it was hard for the coders to interpret their meaning. Thus, visibility had a large effect on how people gestured and the presence of dialogue had a large effect on gesture rate.

Listener needs. Besides mutual visibility and dialogue, Jacobs and Garnham (2006) point out that gesture production may depend on the behavior and needs of the addressee (also see Enfield et al., 2007), and on the type of task that the speaker is performing. They found that narrators produced fewer gestures when they knew that their addressee already knew part of the content of the story they were telling. They also found that speakers produced more gestures when the addressee appeared attentive, than when the addressee appeared inattentive. They therefore concluded that during narrative tasks, gestures are produced primarily for the benefit of the addressee.

Content. Melinger and Levelt (2004) looked at the type of information being represented. They found that speakers who used gestures representing spatial information omitted more critical spatial information from their verbal descriptions than speakers that did not gesture. They showed that some speakers divided information between the gesture and speech modality. This shows that co-speech gestures expressing spatial information can be used communicatively.

Hostetter and Hopkins (2002) have shown that speakers accompanied their narration with more representational gestures (which they term "lexical movements") if they watched an animated cartoon and subsequently were asked "to picture the events they saw in the cartoon in their head and then describe them" (p. 25), than when they read a description of the events in the cartoon and were asked "to picture the words as they had read them on the page and then relate them" (p. 25) while retelling the events. They interpret this as evidence that representational gestures (lexical movements) are produced more frequently when expressing a thought that is encoded spatially, than when expressing a thought that is encoded textually.

Human–machine interaction

Next to the above-described factors that influence gesture production, human–machine interaction is an important factor in our present study as well. Reeves and Nass (1996) state that "people's responses to media are fundamentally social and natural". This is the so-called *media equation* and it applies to everyone. They state that the confusion of mediated life and real life is not rare and inconsequential, and that it cannot be corrected with age, education, or thought. Even though their studies focused on social responses, e.g. empathy, rather than on communicative

behavior, this would suggest that, even if gestures are used to communicate, people would still gesture at computers and other media, since their social responses may underlie their communication.

But, although people do show social responses to media and artificial agents, one can ask whether they do so to the same extent as to human interlocutors, and how exactly this influences their communicative behavior. Aharoni and Fridlund (2007) conducted a study in which participants smiled more and used more silence fillers to a purported human interviewer than to a computer interviewer. In both cases a prerecorded stimulus was used. They found that simply labeling the stimulus as 'human' caused people to be more communicative. In addition, Maes et al. (2007) showed that if speakers assume that their addressee is human, more referential effort will be made than if the speaker assumes the addressee is a computer. Respondents more frequently described more attributes than necessary to identify an object to the presumed human addressee, than towards the presumed computer. These findings suggest that at least in some cases, people are wordier towards human than towards computer addressees.

Present study

We are interested in the effect of the addressee being human or artificial on gesture production, and in whether possible differences in gesturing resulting from this manipulation are informative to naïve observers. This is because the different functional roles that gesture may serve imply different predictions on how people would gesture towards an artificial addressee, and place different requirements on addressees' sensitivity to differences in gesturing. We will first describe our study from the production viewpoint and then our perception study.

If gesturing is mostly a for-speaker process, either facilitating language production or supporting cognition in another way, then with a similar task, we would expect speakers to gesture in the same way, regardless of the addressee. On the other hand, if gestures are produced to communicate or if gesturing is tied to other aspects of human dialogue, then the addressee being human or artificial may very well influence gesture production. Therefore, we compared a condition in which there was a human addressee with a condition in which there was an artificial addressee, keeping other factors as similar as possible.

For this we have made use of computer mediation. We created a situation similar to one-way video conferencing. A speaker was filmed and was told that an addressee was watching the recordings live, in another room. Throughout this paper we refer to this condition as the 'Web cam condition'. In this condition there was one-way visibility, no physical co-presence, no dialogue, but the speaker believed there was a human addressee. In a very similar condition, the speaker was

told instead that the audiovisual signal of the camera went to an audiovisual summarizer that was located in another room. This condition has similar one-way 'visibility' and, as in the Web cam condition, there was no physical co-presence and no dialogue, yet this time the speaker believed there was an artificial addressee.

In both of these settings, participants were asked whether they understood whom they were addressing, before they started their narration. Only if this was clear to them did the experiment proceed. This is different from the tape-recorder condition in the experiment by Bavelas et al. (2008), in which participants were excluded if they had imagined an addressee. Thus, in their tape-recorder condition the addressee was absent entirely rather than artificial. With this design we have also been able to separate the effects of being visible to an (artificial) addressee from the effect of dialogue, since we have been able to introduce a condition in which the speaker could be seen by another person, yet there was no possibility of dialogue.

To control for the effects of physical co-presence and mutual visibility, which are absent in both the condition with the artificial addressee and the condition with a human addressee in another room, we have included two more conditions in our design. These were the conditions used in Alibali et al. (2001): face-to-face communication, in which there is a human addressee, physical co-presence and mutual visibility, and a condition in which speaker and addressee are in the same room, but separated by a wooden screen. Although both of these conditions enable dialogue, we prevented true dialogue from happening by instructing addressees not to interrupt the speaker, but to act naturally otherwise. Thus, addressees were looking at the speaker and gave occasional non-verbal feedback, but they tried to avoid speaking themselves.

For our production study, we asked speakers to retell an animated cartoon in which there are many actions involving direction and moving from one location to another. According to Hostetter and Hopkins (2002), this should lead speakers to produce many representational gestures. And based on the results found by Melinger and Levelt (2004), we would expect speakers to use gestures that express spatial information communicatively in this narration task. Content was always said to be new to the addressee, and as explained above, addressees were instructed not to interrupt the speaker. This was in order to minimize the effects found by Jacobs and Garnham (2006).

Based on the results by Aharoni and Fridlund (2007) as well as the results found by Maes et al. (2007), and the assumption that gesturing bears some communicative function, we would expect participants to produce more gestures in our conditions with a human addressee, than in our condition with an artificial addressee. In addition, based on previous results with imagistic gestures (Bavelas et al., 2008) and pointing gestures (Bangerter & Chevalley, 2007) we would expect

representational gestures to be larger in conditions in which they have communicative potential.

As mentioned in the introduction, we think it unlikely that speakers would put communicative effort into the gesture modality if they never use this modality as a source of information. In addition, a possible difference between gesturing to a human or to an artificial addressee cannot play a significant role in interaction if addressees are ignorant to this difference. We therefore also conducted a perception study, in which we asked participants to judge whether a speaker was talking to a human or to an artificial addressee, based on movie clips from our production study. These clips were played without sound and different conditions included or excluded the hands and face of the speaker.

Production study

Method

Design. As outlined in the previous section under 'Present study', we have used a between subjects design with four conditions. A schematic overview of this design can be found in Figure 1. We are mainly interested in the effect of the addressee being human or artificial, which is the only difference between our Computer condition (1) and Web cam condition (2). In both conditions the speaker does not receive any feedback from the addressee.

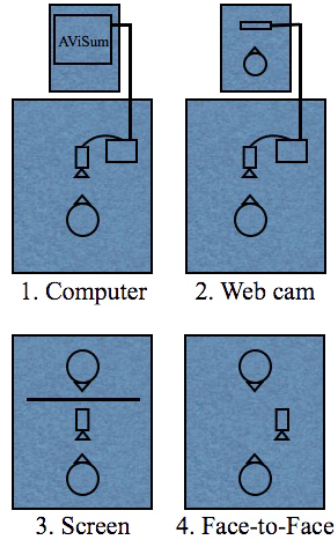

Figure 1. Design.

So that we may see how closely communication through web cam resembles face-to-face communication, we have added a Screen (3) and Face-to-Face (4) condition. Addressees in these conditions were instructed not to interrupt the speaker, but to act naturally otherwise, in order to enhance similarity with the aforementioned conditions. In contrast to the Web cam condition, in both of these conditions the addressee was seated in the same room as the speaker. In the Face-to-Face condition there is mutual visibility as well. If either physical co-presence or seeing the addressee plays a critical role in performing the narration task, then this should result in notable differences between the Web cam, Face-to-Face, and Screen condition. By comparing these three conditions, we get an idea of how closely the Web cam condition resembles a condition with a true and physically present human addressee.

Participants. 43 participants volunteered as narrators for this study. We excluded three participants, because they either were suspicious about the experimental setup or ignored the instructions. The remaining 40 participants (10 male, 30 female) were between the age of 17 and 48 (M = 23, median 19). They were all native speakers of Dutch. None of the participants objected to being recorded, and all of them consented to their data being used for research and educational purposes. There were 11 participants in the Computer condition (condition 1), 10 in the Web cam condition (condition 2), 9 in the Screen condition (condition 3), and 10 in the Face-to-Face (FtF) condition (condition 4). The listeners in the Screen and FtF condition were confederates.

Procedure. We randomly assigned participants to one of four conditions. Narrators first read the instructions (see below for more detail) and could ask any questions they had on the task. The instructions focused on the task of the addressee, namely to summarize the speaker's narration. This way we suggested that the study was on summarizing. Speakers were explicitly asked not to summarize themselves, but to just retell the story. They then watched a seven minute animated cartoon called "Canary Row", which we chose because it has proven to elicit gestures in several other studies, such as McNeill (1992) and Alibali et al. (2001). After being seated in front of the camera, in condition 1 and 2 the experimenter asked whether the participant had understood whom they were going to talk to, and paraphrased their answer if it was correct and elaborated on it if it was incomplete. In condition 3 and 4 the experimenter repeated that the speaker was not to address the camera, but the other participant.

In condition 1, the written instructions said that the signal of the camera was sent to a beta version of an audiovisual summarizer (AViSum) that was located in another building on campus, and which would produce a summary of their narration afterwards. It was emphasized that the system could process both auditory

and visual information. A fake phone call was made by the experimenter to check whether the signal was received well, and whether the system was ready for use. In reality, there was no such computer system. However, it is not inconceivable that such a system could exist. Dupont and Luyten (2000), for example, describe a speech recognition system that uses both acoustic and visual speech information, and McCowan et al. (2005) describe how automatic analysis of meetings can benefit from information from the visual modality.

In condition 2, the instructions said that the camera was used as a web cam, and that another participant was watching the speaker in another campus building, with the purpose of summarizing their narration afterwards. The experimenter pretended to set up a one-way videoconference with a presumed experimenter in the other building, and then made a fake phone call to check whether the image and sound were received well and whether they were ready to begin. In reality, there was no other participant watching.

In condition 3, two students came to the lab, one of which was a confederate. The experimenter pretended to randomly assign the roles of speaker and listener, but always assigned the true participant the role of speaker. After the participant had watched the animated cartoon, narrator and addressee were allowed to ask any questions they had about the task. A wooden screen separated them, such that they could not see each other during the story telling. The narrators' instructions stated that the addressee had to summarize the story afterwards, and that they were videotaped with the purpose of comparing the addressee's summary to their narration. We instructed addressees not to interrupt the narrator, but to act naturally otherwise. Occasionally, there was some auditory feedback (laughs, occasional uh-huh's). Condition 4 was similar to condition 3, except that narrators retold the story in a face-to-face situation, thus without the screen in between narrator and addressee.

In each condition, participants were videotaped using a digital video camera. They were seated in front of the camera. The camera position was such that the entire upper part of the body was visible, including the upper legs. In all conditions, the narrator could look at snapshots of each of the episodes of the cartoon that hung either on the wall or on the screen in front of them. This was done in order to aid memory, and to facilitate more structured, and more comparable stories.

After retelling the cartoon, in condition 3 and 4 the experimenter first took the addressee to another room, supposedly to write the summary. Narrators then completed a questionnaire, which included questions on how they had experienced the conversation and whether they had believed the experimental setup. We fully debriefed all participants and asked their consent to use the recordings. The experimenter also asked whether the participants had believed the experimental setup and whether they had suspected any deceit.

Transcribing and Coding. The first author transcribed each narration from the videotape. Repairs, repeated words, false starts, and filled pauses were included. The annotation of gestures was done blind to condition and initially by the first author. Difficult cases were resolved by discussion among the authors.

Initially, coding concentrated on movements of the hands. Later on, when coding for gesture size, movements of other body parts were considered, but only if they occurred simultaneously with a hand gesture. We first discriminated between gestures and other movements such as self-adjustment. We then coded gestures according to McNeill (1992, pp. 78–82), but adding *interactive gestures*, as described in Bavelas (1992). Gestures were first coded as representational, beat, or interactive. This first division could largely be made based on the shape of the gesture (see McNeill, 1992, and Bavelas, 1992). Simple, biphasic movements of the hands were labeled as beat rather than interactive (in Bavelas's definition, interactive gestures subsume the category of beats). Subsequently, we further divided representational gestures into imagistic (iconic or metaphoric) and pointing gestures. Our most important criterion for labeling a gesture as a pointing gesture was the shape of the hand, which should have one or more fingers extended as an index. In addition, we have judged for each of those gestures whether it seemed to only express information on location or direction, or whether it additionally seemed to express significant information about manner or path. If the latter was the case, the gesture was counted both as imagistic and pointing gesture. So all representational gestures that were not just pointing gestures were counted as imagistic gestures.

In a separate round of gesture coding, we coded for gesture size. Gestures that were produced using only the fingers received a score of 1. If the wrist was moved significantly the gesture received a score of 2. Gestures that also involved significant movement of the elbow or lower arm received a score of 3, and gestures in which the upper arm was also used in a meaningful way, or that involved movement of the shoulder received a score of 4.

Statistical Analysis. For all tests for significance we have used univariate analysis of variance (ANOVA), with condition as the fixed factor (with levels: computer, web cam, screen and face-to-face) and a significance threshold of .05. For pairwise comparisons we have used the least significance difference test (Fisher 1951).

Results

Gesture Rate. Condition had a significant effect on the number of gestures produced per 100 words, $F(3,36) = 6.269$, $p < .01$, $\eta_p^2 = .343$, see Figure 2. Gestures were significantly less frequent in the Computer condition ($M = .64$, $SD = .84$) than in the Web cam ($M = 3.8$, $SD = 4,3$), Screen ($M = 3.7$, $SD = 1.9$), and Face-to-Face

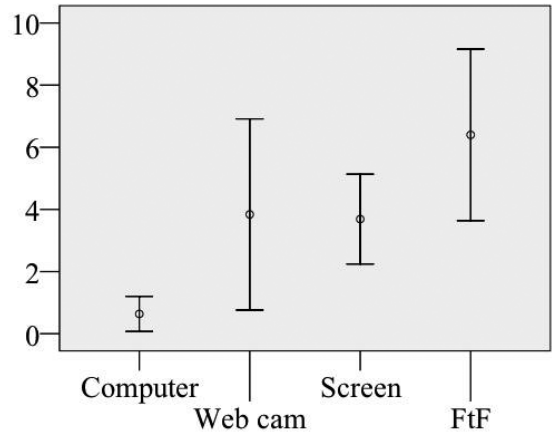

Figure 2. Means and 95% confidence intervals for the average number of hand gestures produced per 100 words, per condition.

(FtF) condition (M = 6.4, SD = 3.9). The differences between the mean gesture rates in the Web cam, Screen, and FtF conditions were not significant.

However, when performing the analysis with gestures per second rather than per word, $F(3, 36) = 7.044$, $p < .001$, $\eta_p^2 = .370$, we found that gestures were reliably more frequent in the FtF condition (M = .22, SD = .13), than in the Screen (M = .12, SD = .06) and Web cam (M = .12, SD = .14) condition. On this analysis also, significantly fewer gestures were produced in the Computer condition (M = .017, SD = .023) than in any of the other three conditions.

Four of the eleven participants in the Computer condition did not produce any gestures. In the other conditions there were no participants that did not gesture at all.

Gesture rate and type. We also found a significant effect of condition on representational gestures per 100 words, $F(3, 36) = 5.658$, $p < .01$, $\eta_p^2 = .320$, see Figure 3. Representational gestures were produced at a reliably lower rate in the Computer condition (M = .37, SD = .55) than in the Web cam (M = 2.9, SD = 3.5) and Face-to-Face condition (M = 4.8, SD = 3.2). There was a trend towards significance for the difference between the Computer and Screen condition (M = 2.4, SD = 1.4), $p = .08$. In the Screen condition, reliably fewer representational gestures were produced than in the FtF condition.

For non-representational gestures per 100 words, we found a significant effect of condition as well, $F(3, 36) = 4.745$, $p < 0.01$, $\eta_p^2 = .283$, see Figure 4. Non-representational gestures were produced at a significantly lower rate in the Computer condition (M = .26, SD = .43) than in the Screen (M = 1.3, SD = .78), and FtF condition (M = 1.6, SD = 1.2) condition. There was a trend towards significance

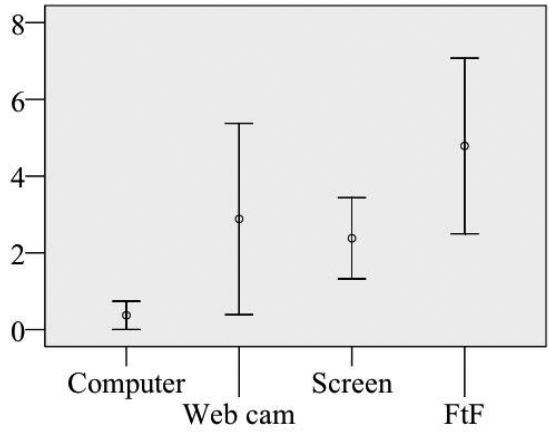

Figure 3. Means and 95% confidence intervals for the average number of representational gestures produced per 100 words, per condition.

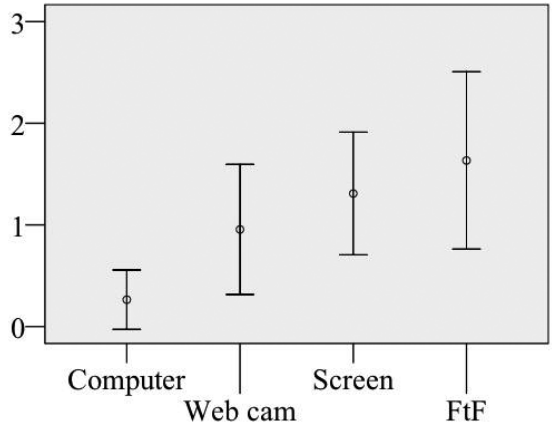

Figure 4. Means and 95% confidence intervals for the average number of non-representational gestures produced per 100 words, per condition.

($p = .08$) for the difference between the Computer and Web cam condition (M = .96, SD = .90). In all conditions, representational gestures occurred more frequently than non-representational gestures.

For imagistic gestures, condition had a significant effect on the mean gesture rate, $F(3, 36) = 5.005$, $p < .01$, $\eta_p^2 = .294$. In the Computer condition (M = .37, SD = .55), imagistic gestures were produced significantly less frequently than in the Web cam (M = 2.5, SD = 3.0) and FtF condition (M = 4.0, SD = 2.9). There was a trend towards significance for the difference between the Screen (M = 2.1, SD = 1.2) and FtF condition ($p = .06$).

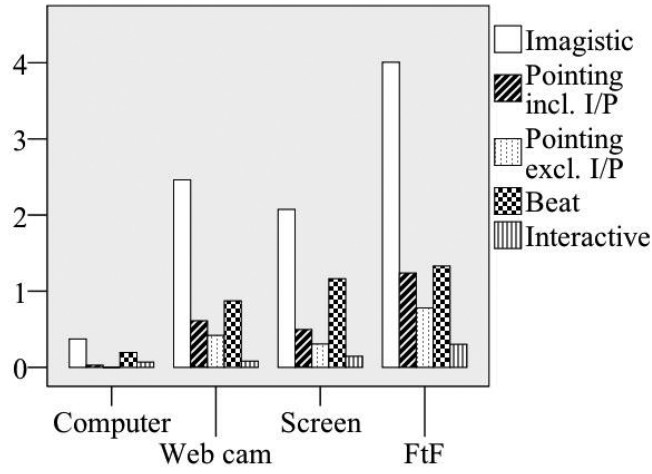

Figure 5. Mean number of gestures produced per 100 words per gesture type (Imagistic; Pointing gestures including imagistic/ pointing gestures; Pointing gestures excluding imagistic/pointing gestures; Beat; Interactive), per condition.

Only one pointing gesture was produced in the Computer condition. This was a combined imagistic/ pointing gesture. Our ANOVA for pointing gestures per 100 words, $F(3, 36) = 4.82$, $p < .01$, $\eta_p^2 = .287$, showed a significant difference between the Screen ($M = .50$, $SD = .65$) and FtF condition ($M = 1.2$, $SD = .80$). There was a trend towards significance ($p = .06$) for the difference between the FtF and Web cam condition ($M = .61$, $SD = 1.0$). The Computer condition ($M = .029$, $SD = .22$) differed significantly from the FtF condition and there was a trend towards significance for the difference between the Computer and Web cam condition, $p = .08$. When combined imagistic/ pointing gestures were excluded from the analysis, we found similar results. Figure 5 shows the mean number of gestures per 100 words for the different gesture types. In the first bar for pointing gestures, pointing gestures that also seemed to convey significant information on manner or path (imagistic/ pointing gestures) are included, in the second they are not.

Gesture Size. Using the coding system described in the previous section, we computed a score that represented the average size of a gesture for each participant. For each participant, we took the sum of the scores of all gestures and divided this sum by the number of gestures produced by that participant. Although overall gesture size did not differ significantly across conditions, $F(3, 32) = 1.341$, $p = .28$, there was, nevertheless, a tendency for gestures to be larger in the conditions where speakers thought that the addressee could see them. Thus, gestures were smallest in the Computer condition ($M = 2.0$, $SD = .80$), followed by the Screen ($M = 2.2$, $SD = .51$), Web cam ($M = 2.4$, $SD = .65$), and FtF condition ($M = 2.6$, $SD = .38$).

When the Computer condition and the FtF condition were compared, however, this did approach significance ($p = .07$).

The proportions of large and small gestures differed across conditions, as can be seen in Figure 6a. Condition had a significant effect on the percentage of gestures that involved shoulder movement, $F(3,32) = 4.039$, $p < .05$, $\eta_p^2 = .275$, see Figure 6b. These gestures made up a significantly larger portion of the total number of gestures in the Web cam condition (M = .16, SD = .14) than in the Computer (M = .029, SD = .076), and Screen condition (M = .014, SD = .031). We found a tendency towards significance ($p = .08$) for the difference between the FtF (M = .097, SD = .11) and Screen condition.

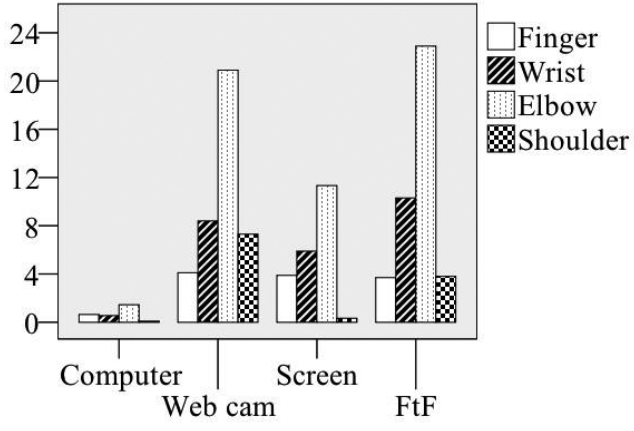

Figure 6a. Mean number of gestures produced per size (Finger, Wrist, Elbow, Shoulder), per condition.

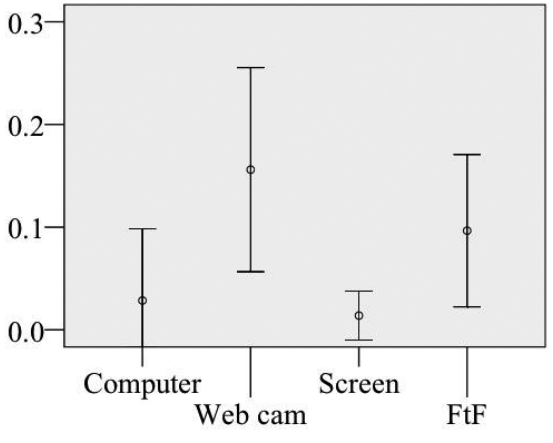

Figure 6b. Means and 95% confidence intervals for the average proportion of gestures that involved movement of the shoulder, per condition.

Gesture size and type. For representational gestures, overall gesture size was very similar across conditions, ranging from M = 2.7 in the Screen condition, to M = 3.0 in the FtF condition, F(3, 27) = 1.084, $p = .37$. We found no significant main effect of condition on the size of imagistic, F(3, 27) = 1.084, $p = .37$, or pointing gestures, F(2, 17) = 2.434, $p = .12$. However, for pointing gestures, gesture size was significantly smaller in the Screen condition (M = 1.7, SD = .82) than in the FtF condition (M = 2.8, SD = .68). Combined imagistic/ pointing gestures were counted as imagistic in this analysis. We found no significant main effect of condition on the size of non-representational gestures, F(3, 31) = 1.856, $p = .16$. No significant differences in the size of interactive gestures, F(3, 14) = .397, $p = .76$, and beats, F (3, 31) = 1.422, $p = .26$, were found either. But post hoc analysis showed that non-representational gestures were significantly smaller in the Computer (M = 1.3, SD = .47) than in the FtF condition (M = 1.9, SD = .36).

Figure 7 gives an overview of the average size scores for the different gesture types for each condition. It must be noted however that some means are derived from very few data points, since some types of gesture were produced by only very few participants in some conditions. Figure 8 gives an overview of the average (over participants) number of gestures produced per gesture type in each condition, and can help to interpret Figure 7.

Number of words. Condition had a significant effect on the total number of words used by participants, F(3, 36) = 4.261, $p < .05$, $\eta_p^2 = .262$, see Figure 9. In the Web cam condition (M = 842, SD = 374), significantly more words were used than in the Computer (M = 473, SD = 131) and FtF condition (M = 595, SD = 268).

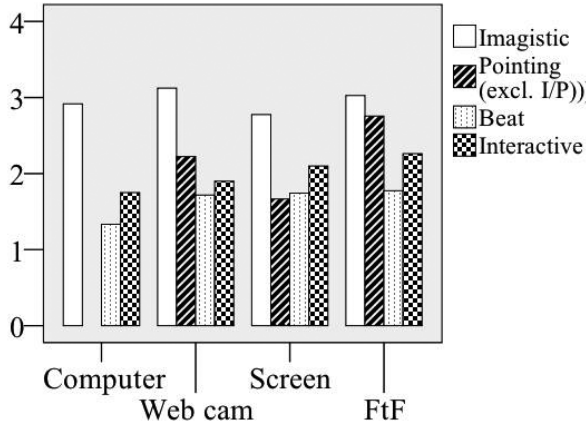

Figure 7. Average size score (1 = Finger, 2 = Wrist, 3 = Elbow, 4 = Shoulder) of gestures produced per gesture type (Imagistic, Pointing, Beat, Interactive), per condition.

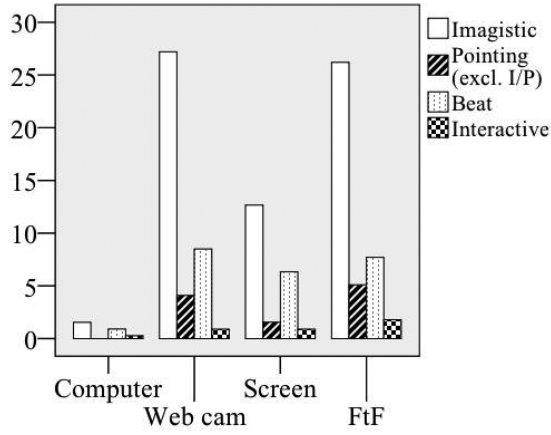

Figure 8. Average number of gestures produced per gesture type, per condition.

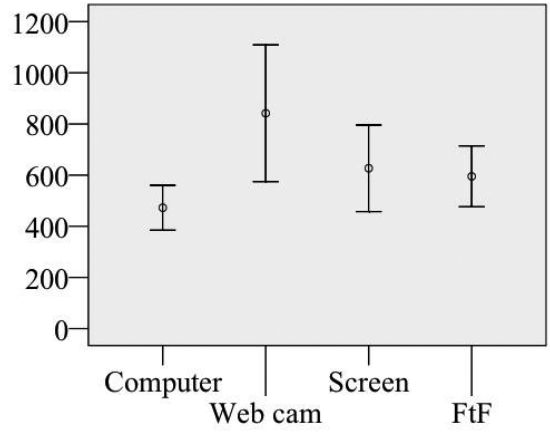

Figure 9. Means and 95% confidence intervals for the total number of words spoken, per condition.

Speech rate. We also found a significant effect of condition on the number of words spoken per second, $F(3, 36) = 4.916$, $p < .01$, $\eta_p^2 = .291$, see Figure 10. Speech was slower in the Computer condition ($M = 2.6$, $SD = .24$) than in the Screen ($M = 3.3$, $SD = .18$) and FtF condition ($M = 3.3$, $SD = .49$). There was a trend towards significance for the difference between the Computer and Web cam condition ($M = 3.0$, $SD = .14$), $p = .07$. Speech was faster when interlocutors were physically co-present, $F(1, 36) = 9.515$, $p < 0.01$, $\eta_p^2 = .209$.

Filled pauses. No significant main effect of condition on the number of filled pauses (i.e. uhs) per word was found, $F(3,36) = 1.816$, $p = .162$. However, post hoc tests

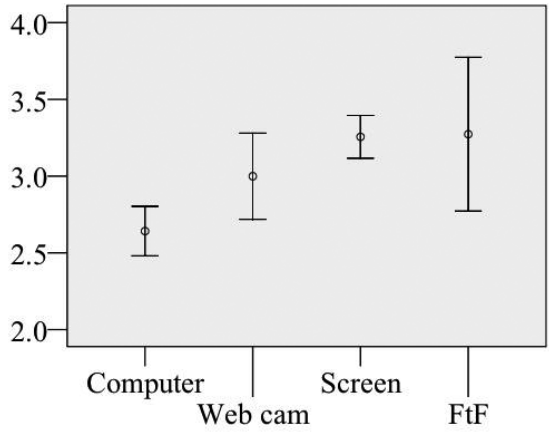

Figure 10. Means and 95% confidence intervals for the average number of words spoken per second, per condition.

showed that filled pauses were more frequent in the Web cam (M = .096, SD = .043) than in the FtF condition (M = .057, SD = .036).

Discussion

Participants who thought that they were talking to an audiovisual summarizer produced fewer gestures than participants who thought they were talking to a human addressee, regardless of whether the addressee was in the same room or not and whether or not there was mutual visibility. Also, gestures produced by participants who believed that they were talking to the computer system were more frequently small (not involving shoulder movement) than were the gestures produced by participants who thought they were talking to a human addressee through the web cam. So the (presumed) nature of the addressee, either human or artificial, clearly influenced gesturing.

The only difference between the Computer and Web cam condition was whether participants were told they were speaking to an audiovisual summarizer, or to another participant. Both were said to be in another room, so in both conditions the participant was narrating in front of a camera, without seeing or receiving any feedback from the addressee. Therefore, the difference in gesture rate and gesture size that we found between these two conditions can only result from speakers' mental representations of the addressee and this representation must include whether the addressee is artificial or not.

More words were used in the Web cam condition than in the Computer condition. Participants also spoke a little more slowly when they thought they were interacting with the computer system. Part of the difference in gesturing that we

found between these two conditions could therefore result from differences in verbal behavior, rather than directly from differences in speakers' knowledge of the addressee's nature. However, the Web cam condition rather than the computer condition is the atypical one when looking at the number of words. The number of words used in the Computer condition did not differ significantly from the number of words used in the Screen and Face-to-face condition, whereas the difference in gesture rate between the Computer condition and these two conditions is striking. Descriptions in the Computer condition were generally detailed and elaborate, just like in the other conditions. We therefore think it unlikely that possible differences in the verbal behavior are the only source of the differences in the gestural behavior that we found. In addition, it would be hard for such a theory to explain why pointing gestures were almost completely absent in the Computer condition, while the same spatial content had to be expressed. Rather, we think that both the verbal and gestural modality were affected by the addressee being an artificial system or a human participant in another room.

Is the comparison between our Web cam and Computer condition a valid way to compare human–human to human–machine communication? As can be seen from Figure 8 and 6a, the gestural behavior of participants in the Web cam condition was very similar to that of participants in the FtF condition. Similar patterns can be observed for the proportions of different gesture types and sizes. When looking at the average gesture rate (Figures 2 and 5), the Web cam condition is more similar to the Screen condition. This indicates that seeing the addressee, or the possibility of dialogue, may play as big a role as being seen by the addressee. The comparison between the Computer and Screen condition shows that the very low gesture rate in the Computer condition does not just result from speakers not being visible to a human addressee. Neither can it be fully explained by the factor of physical co-presence, since the Computer condition differed significantly from both the Web cam and the FtF condition. It thus seems that our design was indeed able to capture the difference between human–human and human–machine communication we were interested in.

The effect of mutual visibility on gesture production was replicated for the number of representational gestures per word, the number of pointing gestures per word, and the size of pointing gestures. For these variables we found significant differences between the Screen and FtF condition, as did earlier studies (i.e., Alibali et al., 2001; Bangerter & Chevalley, 2007).

People behaved very differently towards the artificial system as compared to how they behaved towards people. This is contrary to what would have been expected by Reeves and Nass (1966), for example, who suppose that people behave toward 'media' as they would towards a real person. Since fewer and relatively fewer large gestures were produced and participants did not use more words, it

does not seem that information was mostly transmitted through speech instead of through gestures when talking to the computer. At first glance, it seems that less information was transmitted through both modalities. This corroborates well with the idea that people are less communicative when communicating to computers, as has been suggested by Aharoni and Freedlund (2007) and Maes et al. (2007). It would be interesting to do a comparative analysis of the verbal discourse to arrive at more clarity in this.

The differences we found in gesturing in different communicative settings can be explained by the idea that people make gestures for the benefit of their addressees. But as explained under 'Present Study', we would find this explanation less believable if addressees are not sensitive to such differences. To test whether they are, we conducted two perception experiments, which will be described in the next section.

Perception study

It has been shown that addressees are able to process information from gestures (Beattie & Shovelton, 1999, 2001; Goldin-Meadow & Sandhofer, 1999). However, in these studies information was directly related to the message a speaker was trying to convey, rather than to the communicative setting that a speaker was in. Chawla and Krauss (1994) have found that observers could discriminate better than chance between spontaneous and rehearsed speech, both based on audio and audio-visual presentations.[2] However, it remained unclear what cues observers had used in making their judgments.

With this study we want to determine whether observers are sensitive to differences in gesture production that result from differences in the communicative setting, especially the difference between addressing a human or an artificial addressee. At the same time, this perception study can be seen as a way to independently verify the gesture coding in our production experiment.

Experiment 1

In this experiment we asked observers to watch movie clips that were taken either from a setting with an artificial addressee (the Computer condition of our production study), or with a human addressee (the Screen condition of our production study). To separate the effect of gesturing from the effects of other visual cues, we measured the relative contributions of seeing the face and seeing the upper-body (including hands and arms) of the speaker.

Method

Design. We used a between subjects design with three conditions. In condition 1, the 'Whole speaker condition', participants saw video clips in which the speaker's upper-body was fully visible. In condition 2, participants saw video clips in which the speaker's head was covered by a black rectangle (the 'Hands only condition'). And in condition 3, the 'Face only condition', participants saw video clips showing the head of the speaker only. In all conditions, the video clips were played without sound. After each video clip, participants were asked to judge whether the speaker was talking to a human or to an artificial addressee and to state on a binary scale whether they were certain or uncertain about their decision.

Participants. Ninety first and second year students of Tilburg University and Eindhoven Technical University, all native speakers of Dutch, volunteered for this experiment. Most of them received half an hour of course credits for their participation.

Stimuli. For this experiment we used 18 video clips from our production study, 9 of participants in the Screen condition, in which the story of an animated cartoon was retold to another participant (a confederate) who was sitting behind an opaque screen, and 9 from participants in the Computer condition, in which participants retold the same story to a purported audiovisual summarizer. In both of these settings the speaker was seated in front of a camera.

From each video clip, 30 seconds were selected, starting from the point where the speaker began to describe the sixth episode of the cartoon, in which Sylvester builds a seesaw in order to catapult himself up to the window where Tweety sits. This episode was chosen because it is very prone to elicit gestures. For the Whole speaker condition, we used movie clips in which the speaker and all gestures were fully visible. Two different edited versions were then created, one such that everything was covered except the head of the speaker, for the Face only condition, and one in which the head of the speaker was covered by a black rectangle, for the Hands only condition.

Before the actual experiment started, there were two practice trials, for which video clips similar to the ones in the actual experiment were used. They were of a speaker in the Computer condition, and of a speaker in the Web cam condition of the production study.

Procedure. Participants were randomly assigned to one of the three conditions. First, they read a written instruction and could ask the experimenter any questions they had. The instruction explained the task, but only stated that the participant had to indicate whether the speaker was talking to a human addressee or to an audiovisual speech recognition system. Details about the communicative setting,

such as the difference in visibility (the computer could make use of video whereas the human addressee could not see the speaker) or co-presence (the computer was in another room, whereas the human addressee was in the same room) were not mentioned. Participants then did two practice trials, on which they did not receive any feedback. After the practice trials, the experimenter asked them again whether the task was clear and gave further instruction if necessary. Then followed the actual experiment.

Fragments were shown on a computer monitor. Half of the participants watched them in a certain random order, and the other half in reversed order. After each video fragment the screen turned black for seven seconds. On the black screen a sentence was shown in white, stating which fragment the participant was to fill out. This text disappeared after six seconds. The seven second pause was to be used by the participant to fill out on a paper sheet whether the speaker in the previous clip was talking to an audiovisual speech recognition system or to a human addressee, and whether the participant was certain or uncertain about this judgment (binary scale).

After having judged all video clips, participants completed a brief questionnaire asking what features of the stimuli they had used in judging.

Results

Error rate. The error rate refers to the proportion of movie clips that were judged incorrectly by a participant. We found a significant effect of condition on the average error rate, $F(2,87) = 6.680$ $p < .01$, $\eta_p^2 = .133$. The error rate was significantly higher in the Face only condition ($M = .34$, $SD = .12$) than in the condition in which participants could see the speaker entirely ($M = .22$, $SD = .10$), and in the

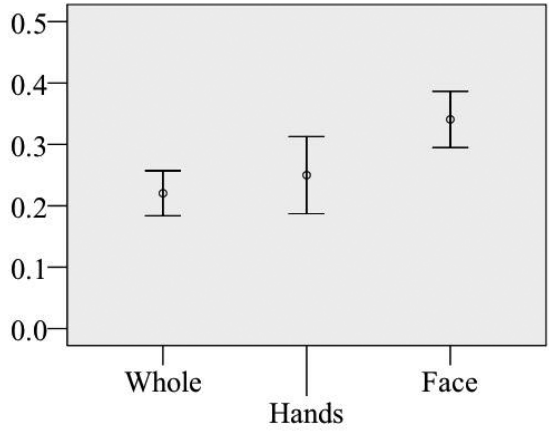

Figure 11. Means and 95% confidence intervals for error rates, per condition.

condition where the face could not be seen (M = .25, SD = .17), see Figure 11. The latter two conditions did not differ significantly. One sample t-tests showed that the error rate was significantly below chance (.5) in all conditions. For the Whole speaker condition: one-sample $t(29) = -15.57$, $p < .0001$, for the Hands only condition: one-sample $t(29) = -8.15$, $p < .0001$, and for the Face only condition: one-sample $t(29) = -7.11$, $p < .0001$.

We also found significant correlations between the number of gestures in our coding of the fragments and the number of participants who thought the speaker was talking to a human addressee, $r(28) = .88$, $p < .01$, for the condition in which the entire speaker was visible and $r(28) = .81$, $p < .01$, for the Hands only condition.

Discussion

The results of experiment 1 suggest that hand gestures are an important cue when judging whether a speaker is addressing a human addressee or a computer system. Participants could make this judgment reliably better than chance, even when they only saw the hands and upper-body of the speaker (without the face), and could not hear the speaker. They had the correct intuition that more hand gestures were produced towards a human, than towards the artificial addressee. The difference in gesturing that we found by analysis of the movie clips from our production experiment thus was confirmed by untrained observers, who could see parts of the movie clips only once.

For this first perception experiment, we have compared movie clips from the Computer condition to movie clips from our Screen condition of our production experiment, rather than from our Web cam condition. Gesture rate, and overall gesture size did not differ significantly between these two conditions, although more very large gestures were produced in the Web cam condition. Also, in neither of these conditions did speakers receive visual feedback from the addressee. There was occasional auditory feedback in the Screen condition, but this was so rare that we trust it not to have had a major influence on our results, which is also indicated by the non-differing gesture rates. Nevertheless, one could argue that the differences that observers in the perception experiment made use of, resulted from a difference between the Computer and Screen condition of our production study other than the difference in the nature of the addressee. We therefore did a control experiment, in which movie clips of speakers from the Computer and Web cam condition were compared, to see whether participants could still reliably judge the nature of the addressee.

Experiment 2

Experiment 2 was similar to experiment 1, but this time we used movie clips from the Computer and Web cam condition of our production study. There was only one condition, in which participants could see the entire upper-body of the speaker. The instruction asked participants to judge whether a speaker was talking to an audiovisual speech recognition system in another room, or to a human addressee, who was watching them live on video from another room. Movie clips were played without sound, and were projected life-size onto a wall. Sixty Master students of Tilburg University, all native speakers of Dutch, volunteered to participate in this experiment.

Results

A one-sample t-test showed that the error rate ($M = .33$, $SD = .08$) was significantly below chance (.5), one-sample $t(59) = -16.87$, $p < .0001$. The correlation between the number of hand gestures in our annotation and the number of participants that thought a speaker was talking to a human addressee was $r(58) = 0.81$, $p < .001$.

Discussion

The results of our perception experiments clearly confirm that there are differences in gesture production when talking to a human addressee or to a computer system (even though the human addressees in the production experiment could not always see the speaker). More importantly, they show that observers are sensitive to these differences and have an intuition about how speakers gesture when talking to a human addressee or to an artificial system. When asked afterwards to explain the basis of their judgments, most participants answered that they thought more gestures would be produced when talking to a human addressee, which is indeed the case.

Many participants also made comments on facial expressions. They expected speakers to be more vivid towards human addressees. Though gestures were the better cue for judging movie clips in experiment 1, we do not conclude that information from the face is less relevant to addressees. We did not inform viewers in experiment 1 that speakers could not see their addressee, or be seen by their addressee. Therefore, information from the face may have been misleading. Also, mutual visibility may influence facial expressions more than it does gesturing.

Even though in both studies participants performed better than chance, the error rate was lower in experiment 1, than it was in experiment 2. We think this may have to do with differences between speakers. Individual differences in

gesture rate were relatively large among speakers from the Web cam condition. Apparently, some speakers matched the observers' expectations better than others. Participants expected speakers in the Web cam condition to gesture more than speakers from the Computer condition, but for some speakers this difference was quite small. This may have to do with the selection of the fragments from the speakers' narrations. We chose an episode in which relatively many gestures were produced, which causes there to be relatively many gestures especially in the Computer condition, in which usually only a few gestures occurred throughout the entire narration. In addition, some speakers, especially in the Web cam condition, may have had more difficulty imagining their addressee than others.

General discussion

Our production study has shown that just the speaker's idea of the nature of the addressee can be enough to influence gesture rate, the type of gestures produced, and the size of the produced gestures. Speakers gestured a lot more towards human addressees than towards a presumed audiovisual summarizer, they did not make pointing gestures towards the artificial addressee, and gestures that involved movement of the shoulder made up a larger portion of the gestures when talking to a human addressee through web cam than when talking to the artificial system. It has been shown that people can largely refrain from gesturing, and do so spontaneously when asked to retell a story to a computer system. We can therefore conclude that gesture production is not a fully automated process and that it is tightly related to the addressee.

Why would it be that people hardly produce gestures towards an audiovisual summarizer? One reason may be that information in gestures is largely redundant with information in speech. It could be that people do not expect a computer system to need such redundant information. Or perhaps gestures are not symbolic enough in nature, but rather relate to knowledge of the world too directly for speakers to expect the computer to benefit from them. Another reason may be that speakers did not feel the need to accommodate the artificial system as much as a human addressee. Branigan et al. (in press) found that speakers adapt less to an artificial addressee provided that it does not give feedback (also see Maes et al., 2007). This may have caused speakers to be less informative in the gestural modality, but also they may have felt free to speak as slowly as they needed towards the artificial system, thereby not needing gestures to "organize rich spatio-motoric information" (Kita, 2000, p. 163) or to facilitate word retrieval (Krauss, 1998). It would be interesting to measure the effect of time pressure on gesturing to test these hypotheses.

From the perspective that gestures are intended communicatively, the question remains open why the difference in gesturing is not more dramatic when people can or cannot be seen by their addressee. This may have had to do with the relative unresponsiveness of our addressees. Another possibility is that it was difficult for participants to apply their knowledge that the addressee could or could not see them. It has been shown for example that people do not always make optimal use of their knowledge of what the addressee can and cannot see when interpreting referential expressions (Keysar et al., 2003). The small difference between the gesture rates in our Screen and Web cam condition somewhat points in this direction. One would expect speakers to gesture more frequently in the Web cam condition, in which they can be seen by their addressee, yet we found very comparable gesture rates for each gesture type in the Web cam and Screen condition.

If speakers indeed had problems applying their knowledge of the addressee, then the difficulty of the narration task may have further contributed to speakers not fully adjusting their behavior to the communicative setting. Most participants had some problems remembering parts of the animated cartoon they were retelling. Both processes: using one's knowledge of the addressee *and* remembering the story of the animated cartoon, may compete for the same cognitive resources. In a follow-up experiment we are manipulating the memory demands of the narration task, to observe whether participants adapt their (verbal and non-verbal) language production more to the communicative setting when doing an easier task, or whether they always gesture less when memory demands are lower.

A third possible explanation of our results is that gesturing is foremost a social activity. This social aspect may be a largely automated process that is simply not applicable when interacting with a computer sysytem. This corroborates well with the ideas in Bavelas et al. (1992) and Bavelas et al. (2008) about gestures having an interactive function. But it goes against the idea formulated by Reeves and Nass (1996), that people's responses to media are fundamentally social in nature. However, their studies did not avoid personalizing the computer by, for example, asking questions such as 'Did the computer help you well?', whereas we carefully formulated our instructions without attributing human actions, qualities or intentions to the audiovisual summarizer. The wording of such questions and instructions may influence the way participants think about an artificial system.

It has also been shown by Bavelas et al. (2008) that the difference between gesturing on the phone and in a face-to-face situation is qualitative rather than quantitative. Gesturing on the phone or to a person behind a screen may thus serve a different purpose than does gesturing face-to-face. Still, our study shows that even this type of gesturing has something to do with interpersonal communication, besides the effect of dialogue, and may not be fully automated.

The effects of visibility on different gesture types that we found corroborate well with the results found by Alibali et al. (2001). For representational gestures we found that significantly more gestures were produced in the Face-to-face condition, in which speaker and addressee could see each other, than in the Screen condition in which they could not. This supports the hypothesis that representational gestures can be intended for the addressee. However, we did not find this difference between the Web cam and Screen condition. In the Web cam condition addressees were said to be able to see the speaker, but the speaker could not see the addressee and speaker and addressee were not physically co-present. One or both of these factors may influence the rate of representational gestures produced.

For non-representational gestures, we found no significant difference between the Face-to-face and Screen condition. However, we did find a difference between the Computer condition and the conditions with a human addressee, which may point to a communicative function of these gestures. In both the study by Alibali et al. and our study, large individual differences between speakers were found.

Like Bangerter and Chevalley (2007), we found an effect of visibility on the size of pointing gestures. Pointing gestures were larger in the Face-to-face, than in the Screen condition. We also found that fewer pointing gestures were produced in the Screen condition than in the FtF condition and that no pointing gestures were produced towards the audiovisual summarizer. This supports the idea that pointing gestures are meant communicatively and that their size is relevant for their meaning (also see Enfield et al, 2007).

Our perception study has shown that gestures can be highly informative about the communicative setting that a speaker was in. Even when only seeing a speaker's gestures and not hearing the speaker, viewers could reliably judge whether that speaker had been talking to a human addressee or to an artificial system. This is consistent with a theory that speakers intend their gestures communicatively, as well as with a theory that speakers gesture mostly for themselves.

Conclusion

Whether the addressee is human or artificial can have an important influence on gesture production. People gesture less and produce a smaller proportion of gestures involving shoulder movement when narrating to an audiovisual summarizer, than when narrating to a human addressee. In addition, almost no pointing gestures were produced towards the artificial addressee. Just the speaker's mental representation of the nature of the addressee (either human or artificial) can be sufficient to influence the number and size of the gestures produced. We therefore

conclude that gesture production is not a process that is fully automated in every communicative setting.

Given the size of the difference in gesture production that we found between narrating towards a human and an artificial addressee, it seems unlikely that gestures solely facilitate speech production. Rather, we think that some gestures are intended communicatively. However, part of the difference in gesturing that we found may relate to differences in verbal behavior.

A speaker's gestural behavior can convey information about the communicative setting that the speaker is in. It can reveal whether a speaker is talking to a human addressee or to a computer system. People are able to make this judgment better than chance from watching a speaker's hand gesture behavior alone.

Acknowledgements

We thank Carel van Wijk for his help in the statistical analysis, and Martin Reynaert and Lennard van de Laar for their technical support. We also thank Jan-Peter de Ruiter and Adam Kendon for the helpful and constructive comments on earlier versions of this article. Preliminary versions of this work were presented at ISGS 2007, AVSP 2007, and a workshop and master class on multimodal metaphors in Driebergen, 2007. Many thanks to the audiences and participants for their inspiring comments and questions on this research; in particular, we would like to thank Sotaro Kita, Barbara Tversky, and Alan Cienky.

Notes

1. See McNeill (1992) for more information on different gesture types.
2. Speech was produced by professional actors in their study.

References

Aharoni, Eyal & Alan J. Fridlund (2007). Social reactions toward people vs. computers: How mere labels shape interactions. *Computers in Human Behavior*, 23, 2175–2189.

Alibali, Martha W., Dana C. Heath, & Heather J. Myers (2001). Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen. *Journal of Memory and Language*, 44, 169–188.

Bangerter, Adrian & Eric Chevalley (2007). Pointing and describing in referential communication: When are pointing gestures used to communicate? In Ielka Van der Sluis, Mariët Theune, Ehud Reiter, & Emiel Krahmer (Eds.), *CTIT Proceedings of the Workshop on Multimodal Output Generation (MOG), Aberdeen, Scotland, January 2007*.

Bavelas, Janet, Nicole Chovil, Douglas A. Lawrie, & Allan Wade (1992). Interactive gestures. *Discourse Processes*, 15, 469–489.

Bavelas, Janet, Jennifer Gerwing, Chantelle Sutton, & Danielle Prevost (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58, 495–520.

Beattie, Geoffrey W. & Heather Shovelton (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, 18, 438–462.

Beattie, Geoffrey W. & Heather Shovelton (2002). An experimental investigation of some properties of individual iconic gestures that affect their communicative power. *British Journal of Psychology*, 93 (2), 179–72.

Branigan, Holly P., Martin Pickering, Jamie Pearson, & Janet F. McLean (in press). Linguistic alignment between people and computers, *Journal of Pragmatics.*

Chawla, Purnima & Robert M. Krauss (1994). Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology*, 30, 580–601.

Cohen, Akiba A. (1977). The communicative functions of hand illustrators. *Journal of Communications*, 27 (4), 54–63.

Cohen, Akiba A. & Randall P. Harrison (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology*, 28, 276–279.

De Ruiter, Jan Peter (1998). Gesture and speech production. Unpublished dissertation, University of Nijmegen.

Dupond, Stéphane & Juergen Luettin (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Muldimedia*, 2 (3), 141–151.

Enfield, Nick J., Sotaro Kita, Jan Peter de Ruiter (2007). Primary and secondary pragmatic functions of pointing gestures. *Journal of Pragmatics*, 39, 1722–1741.

Fish, Robert S., Robert E. Kraut, Robert W. Root, & Ronald E. Rice (1992). Evaluating video as a technique for informal communication. In *Proceedings of the SIGCHI conference on human factors in computing systems, Monterey, CA, USA, May 1992.*

Fisher, Ronald A. (1951). *The design of experiments.* 6th ed. Edinburgh: Oliver & Boyd.

Goldin-Meadow, Susan (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3 (11), 419–429.

Goldin-Meadow, Susan & Catherine Momeni Sandhofer (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, 2, 67–74.

Goldin-Meadow, Susan, Howard Nusbaum, Spencer D. Kelly, & Susan Wagner (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, 12 (6), 516–522.

Hadar, Uri (1989). Two types of gesture and their role in speech production. *Journal of Language and Social Psychology*, 8, 221–228.

Hostetter, Autumn. B., Martha W. Alibali, & Sotaro Kita (2007). Does sitting on your hands make you bite your tongue? The effects of gesture inhibition on speech during motor descriptions. In Danielle S. McNamara & Greg Trafton (Eds.), *Proceedings of the 29th annual meeting of the Cognitive Science Society,* 1097–1102. Mawhah, NJ: Erlbaum.

Hostetter, Autumn.B., William D. Hopkins (2002). The effect of thought structure on the production of lexical movements. *Brain and Language*, 82, 22–29.

Jacobs, Naomi, & Alan Garnham (2006). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, 26, 291–303.

Kendon, Adam (1994). Do gestures communicate? A review. *Research on Language and Social Interaction*, 27 (3), 175–200.
Kendon, Adam (2004). *Gesture: visible action as utterance*. Cambridge: Cambridge University Press.
Keysar, Boaz, Shuhong Lin, & Dale J. Barr (2003). Limits on theory of mind use in adults. *Cognition*, 89, 25–41.
Kita, Sotaro (2000). How representational gestures help speaking. In: David McNeill (Ed.), *Language and gesture* (pp. 162–185). Cambridge: Cambridge University Press.
Krauss, Robert M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science*, 7, 54–60.
Maes, Alfons, Pascal Marcelis, & Frank Verheyen (2007). Referential collaboration with computers: do we treat computer addressees like humans. In Monika Schwarz-Friesel, Manfred Consten, & Mareile Knees (Eds.), *Anaphors in text: cognitive, formal and applied approaches to anaphoric reference* (pp. 49–68). Amsterdam: John Benjamins Publishing Company.
McCowan, Iain, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, & Dong Zhang (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (3), 305–317.
McNeill, David (1992). *Hand and mind: what gestures reveal about thought.* Chicago: University of Chicago Press.
Melinger, Alissa & Willem J. M. Levelt (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4 (2), 119–141.
Özyürek, Asli (2002). Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language*, 46, 688–704.
Reeves, Byron & Clifford Nass (1996). *The Media Equation, how people treat computers, television, and new media like real people and places*. New York: Cambridge University Press (CLSI Publications).

Authors' address

Lisette Mol, Emiel Krahmer, Alfons Maes, and Marc Swerts
Faculty of Humanities, Communication and Cognition
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

l.mol@uvt.nl, e.j.krahmer@uvt.nl, maes@uvt.nl, m.g.j.swerts@uvt.nl

About the authors

Lisette Mol obtained a Master's Degree in Artificial Intelligence (cum laude) at the University of Groningen (The Netherlands). She then joined the ACT-R research group of John Anderson at Carnegie Mellon University (Pittsburgh, PA, USA) as a learning member and research programmer. Currently she is working towards her PhD at the Communication and Cognition group of Tilburg University. Her research focuses on the influence of the communicative setting on

speakers' mental representations of the addressee and their language use. In 2007, she presented part of this research at the international conference of the ISGS.

Emiel Krahmer (PhD 1995) is a computational linguist by training, and a full Professor in the Department of Communication and Information Sciences, Faculty of Humanities at Tilburg University. His research is aimed at getting a better understanding of how humans exchange information during communication (both verbally and non-verbally, using speech, gestures and facial expressions), which in turn may help improving the way computers present information to and communicate with human users. Together with Marc Swerts, he recently published an article on beat gestures in *Journal of Memory and Language*.

Alfons Maes (PhD 1991) is a full Professor in the Department of Communication and Information Sciences, Faculty of Humanities at Tilburg University. His work includes research on the way in which humans adapt their referential behavior to specific communicative conditions (like the type of task or addressee). Another area of his research is the design of multimodal documents in different communicative domains, including advertising, health communication, instructive communication, and websites.

Marc Swerts (PhD 1994) is a full Professor in the Department of Communication and Information Sciences of the Faculty of Humanities at Tilburg University. Swerts has worked and published on various topics, related to (1) the evaluation of different components of spoken dialogue systems, (2) the variability in the phonetic structure of spoken Dutch, (3) cross-linguistic analyses of prosodic cues to information structure, and (4) audiovisual correlates of pragmatic functions (which is currently his main research interest). He has served on the editorial board of major journals in the field of language and speech, and has been a visiting researcher at academic and industrial research institutes in the USA, Japan, Sweden, Italy, and Romania.